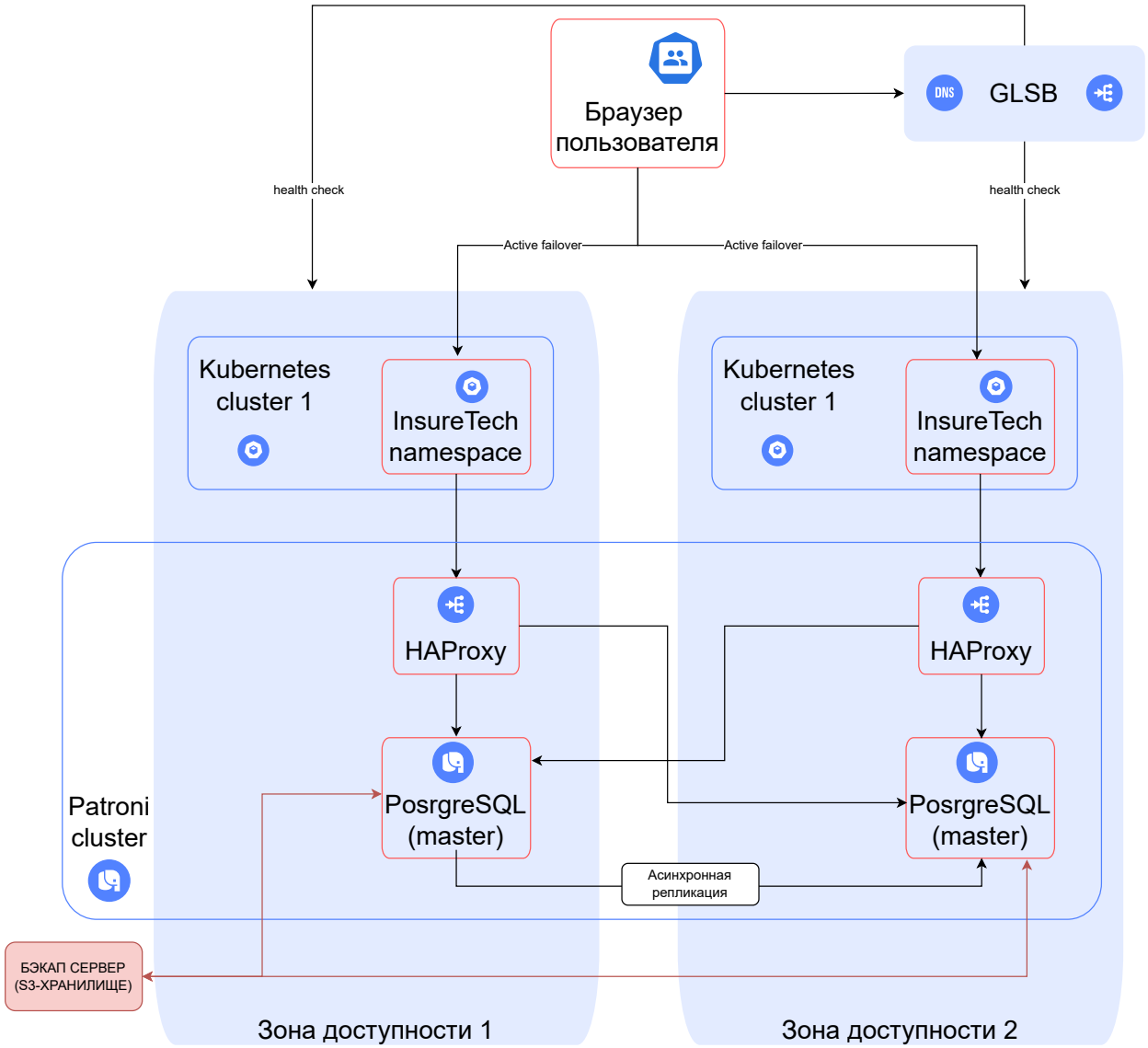


Краткие ответы на поставленные вопросы:

1. Стратегия масштабирования и отказоустойчивости:

Делать упор на горизонтальное масштабирование (POD'ы в Kubernetes). Использовать дополнительные зоны доступности (как минимум 2) для повышения отказоустойчивости. Все сервисы (особенно core-app, client-info, ins-product-aggregator) развертываются в Kubernetes с использованием **Horizontal Pod Autoscaler (HPA)**. Это обеспечит автоматическое добавление подов при росте нагрузки. Для core-app (монолита) — постепенное разделение на микросервисы (например, выделение поиска продуктов и оформления заявок в отдельные сервисы).

2. Деплой приложения в нескольких зонах:

a. Два независимых кластера (K8s) в разных регионах. Это упростит схему фейловера и повысит автономность.

b. Балансировка нагрузки: снаружи — GSLB (Global Server Load Balancer), внутри кластера — Ingress Controller. Health checks настроены как на GSLB (доступность кластера), так и на уровне Kubernetes (готовность POD'ов).

c. Фейловер-стратегия: трафик распределяется между всеми доступными регионами (Active-Active). Для БД обычно Active-Standby. При падении одной зоны GSLB перенаправляет запросы во вторую, БД фейловерится (автоматически или вручную).

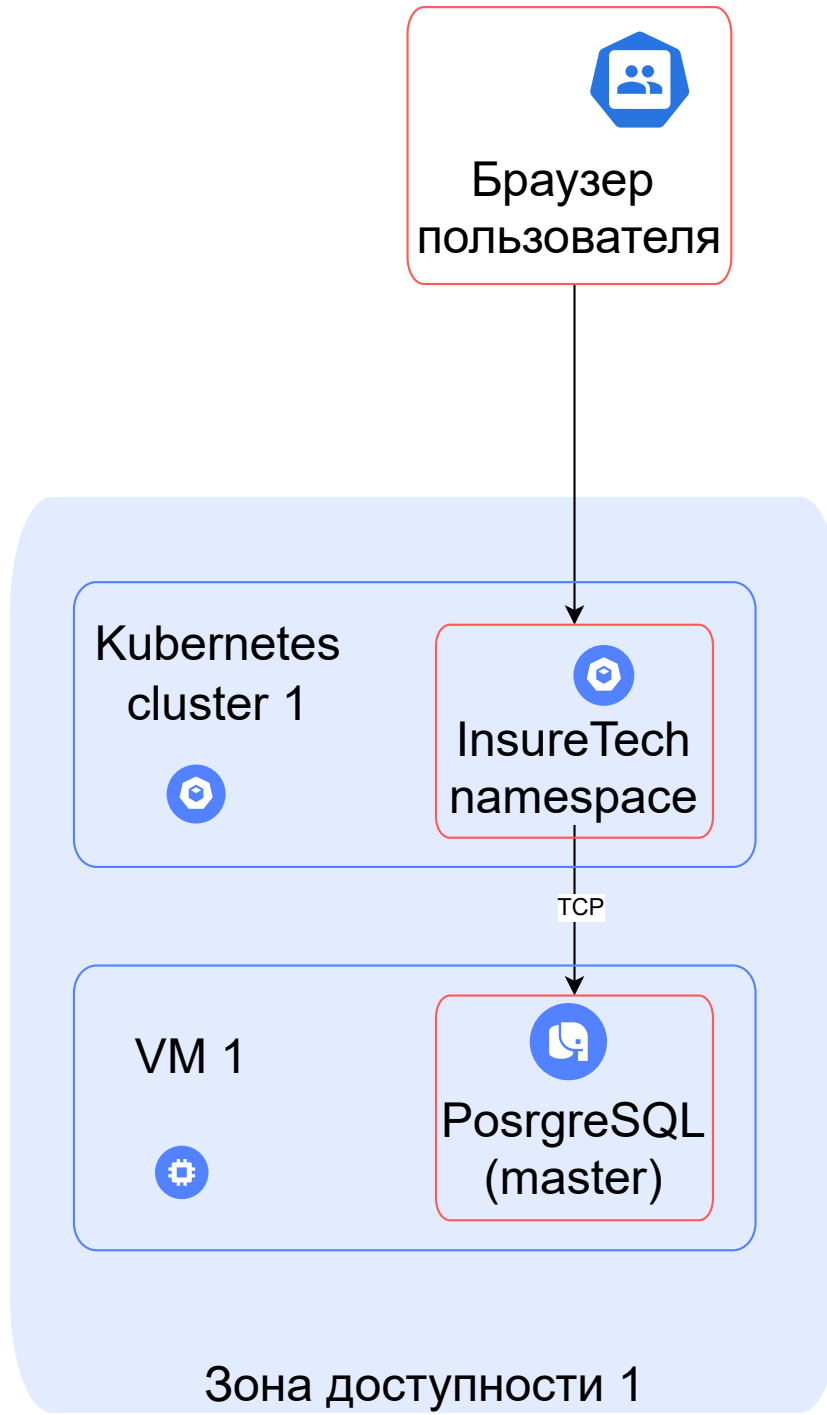
d. Конфигурация БД: PostgreSQL с Patroni обеспечивает автоматическое управление кластером с синхронной репликацией в пределах региона (3 ноды: 1 мастер, 2 синхронные реплики), асинхронной репликацией в другие регионы для минимизации задержек и резервным копированием через ежедневные снапшоты и непрерывную архивацию WAL, позволяя восстановление до точки за 15 минут, при этом бэкапы хранятся в объектном хранилище (S3) с межрегиональной репликацией.

3. Шардирование БД:

С учётом текущего объёма ~50 GB — не требуется. Возможно позже, если объемы данных и нагрузка возрастут кратно.

Итог:

Предложенная архитектура обеспечивает высокую доступность за счёт двух независимых площадок и реплицируемой базы данных, гибкое масштабирование через горизонтальное увеличение POD'ов в Kubernetes, выполнение RTO и RPO (45 мин / 15 мин) за счёт репликации БД и резервных копий, защиту от перегрузки через балансировку нагрузки, единообразное время отклика по регионам РФ (GSLB). В результате система удовлетворяет требованию 24/7, поддерживает SLA 99,9% и готова обрабатывать существенно больший поток запросов (как со стороны B2C, так и B2B).


Браузер пользователя
Kubernetes cluster 1
InsureTech namespace
TCP
VM 1
PosrgreSQL (master)
Зона доступности 1